Our Work:
- Deep learning solutions to integrate the articulatory and emotional features from speech
- Using MTL to capture the relationship between speech articulation and emotional content
- The primary task of predicting lip movements is complemented with two secondary tasks:
  - Emotion recognition
  - Audio: 25 MFCCs
  - 88 eGeMAPS per turn
- Orofacial region:
  - 15 motion capture markers (3x15D)

Rendering Toolkit:
- Xface

Resources
- Corpus:
  - IEMOCAP (1st female subject)
  - Removing idiosyncratic differences
  - Emotion annotations for six emotional categories
  - Three annotators (majority vote)
  - 14 viseme categories
- Background:
  - Lower facial region is modulated by articulatory and emotional cues
  - Modeling these factors can be useful for the generation of expressive lip movements
- Objective Evaluation

Motivation
- Background:
  - Lower facial region is modulated by articulatory and emotional cues
  - Modeling these factors can be useful for the generation of expressive lip movements
- Our Work:
  - Deep learning solutions to integrate the articulatory and emotional features from speech
  - Using MTL to capture the relationship between speech articulation and emotional content
  - The primary task of predicting lip movements is complemented with two secondary tasks:
    - Viseme recognition
    - Emotion recognition
  - Adding auxiliary tasks helps the network to learn more predictive features for orofacial movements

Comparison of MTL and Baselines

Multitask Learning (MTL):
- MTL jointly solves related problems
- Prediction of orofacial movements is the primary task
- Triviseme and emotion recognition are auxiliary tasks
- The auxiliary tasks can be considered as regularizers for the network to learn more and generalizable robust features

\[
\ell = \sum_{i=1}^{N} \lambda^p \ell^p (y^p_i, f (x_i; W^p)) + \sum_{i=1}^{N} \sum_{a \in A} \lambda^a \ell^a (y^a_i, g^a (x_i; W^a))
\]

Discussion
- Using effective regularization in deep learning is important for modeling expressive facial movements
- The secondary tasks were carefully selected to improve the performance of the primary task
- An important strength of our framework is that we can train MTL using datasets with partial information

Future Work
- Modeling idiosyncratic differences between speakers that can be directly added to our models to create personality traits
- Evaluating whether the emotional content conveyed over the orofacial area is preserved in the generated movements

This work was funded by NSF (IIS 1718944)